Guidelines for Working with Chinese UD Treebanks

Language Technologies Institute, Carnegie Mellon University

May 13, 2020

Contents

Chapter 1

Introduction

We hope to provide some guidelines for Chinese UD treebanking, and, for the difficult cases, describe our reasoning for our recommendations. Our guideline includes word segmentation, part-of-speech tagging, and syntactic relations. We choose to focus on the closed class words. We will refer to sentences in our parallel dataset on *The Little Prince*, which can be found here.

Chapter 2

Segmentation

Penn Chinese Treebank has detailed guidelines on how to perform word segmentation in Chinese. See The Segmentation Guidelines for the Penn Chinese Treebank (3.0) [1]. Although Penn Chinese Treebank has a different set of POS tags, the same segmentation scheme can still be used for UD. While applying these guidelines, we made the following adjustments to make them suitable for UD.

- Compound verbs: in general, a lot of verb compounds are treated as one word or "a word with internal structure" in the guidelines. For example, 找到, 写出, 写了出来, etc. In UD, there are syntactic relations (COMPOUND:DIR, COMPOUND:VO, and COMPOUND:VV) designed to deal with compound verbs as separate words. Therefore, we decided to segment these compound verb phrases into separate words.

In addition, these are some of the difficult cases.

- Ordinal numbers, like 第一, are one word.
- 们 (men2) is attached to the noun. So 大人们 is one word.
- Reflexive pronouns, like 他们自己, are one word.
- Reduplicated verbs are one word. For example, 测试测试, 看了看.

Chapter 3

Part-of-Speech Tagging

Here we give a brief overview of the words in each part-of-speech category. Again, we only focus on the closed classes.

3.1 PRON and DET and NUM

Some words can function as both PRON and DET. This includes:

- 这 (this), 这些 (these), 那 (that), and 那些 (those)
- 任何 (any) and 有些 (some)
- 什么 (what), 多少 (how many)

Other words in PRON include:

- Personal pronouns: 我 (I), 他们 (they), 你自己 (yourself)
- Locational pronouns: 这里 (here), 那里 (there)
- Interrogative pronouns: 谁 (who), 哪里 (where)

Other words in DET include:

- Quantifying determiners: 每 (every), 各 (each), 全 (all) , etc.
- Interrogative determiners: 哪 (which), 哪样 (which kind)

3.2 NUM

The class NUM contains all forms of cardinal numbers. On the other hand, ordinal numbers are all tagged ADJ.

3.3 ADP

The adpositions belong to the following three categories. A noun can have both a preposition and a postposition.

3.3.1 Preposition

The common prepositions are 在 (at), 对 (towards), 和 (with), etc.

3.3.2 Postposition

The postpositions are words such as 上 (above), 下 (below), 中 (middle), 前 (in front), etc.

3.3.3 Valence markers

This class includes 把 (ba3) and 被 (bei4). Note that 被 can also be AUX in some cases.

3.4 AUX

According to UD documentation, the auxiliaries fall in the following three categories.

3.4.1 Model auxiliaries

Some these words are: 能 (can), 会 (will), 应该 (should), 有 (you3) (as perfective), 没有 (mei2you3) (as negative perfective). The last two can also appear as verbs.

3.4.2 Aspect markers

The aspect markers are 了 (le5), 着 (zhe5), and 过 (guo4).

3.4.3 Copulae

The word 是 (shi4) is in this category, as well as other less common copulae like 为 (wei2).

3.5 PART

3.5.1 The three de5's

- 的 (de5) always has tag PART, regardless of how it is used.

- 地 (de5), used as an adverbializer, is also in PART.
- Same with 得 (de5), which is used in the construction of complement phrases.

3.5.2 Sentence-final Particles

Some words in this category are 的 (de5), 吗 (ma1), and 吧 (ba5).

3.5.3 Other Particles

According to UD documentation, words like 等等 (deng2deng3) and 所 (suo3) are also tagged as PART.

3.6 CCONJ and SCONJ

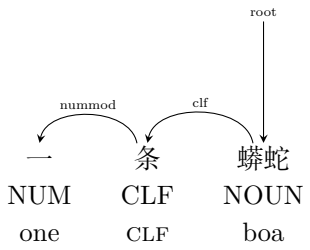
- Coordinating conjunctions are words such as 和 (and), 并且 (in addition), 或 (or), 或者 (or), 但 (but), 但是 (but/however).

  Note that words such as 和 can also function as adpositions.
- Subordinating conjunctions are words such as 如果 (if), 虽然 (although), and 来 (in order to).

Chapter 4

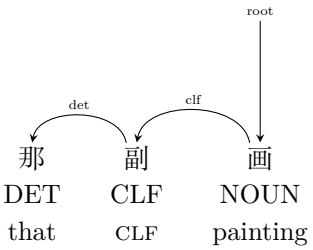
Syntactic Relations

4.1 Classifiers



Tree 1: classifier, from `sent_id=1_2`

The number can be replaced by a determiner, and the structure remains the same.



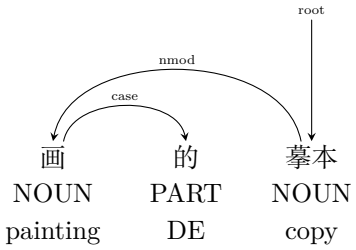
Tree 2: classifier with determiner, from `sent_id=1_3`

4.2 的 (de5)

的 has the following four different uses.

4.2.1 Noun + 的 + Noun

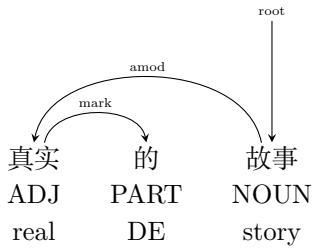
Here 的 is treated as the case marker for the first noun.



Tree 3: noun + 的 + noun, from `sent_id=1_3`

4.2.2 Adjective + 的 + Noun

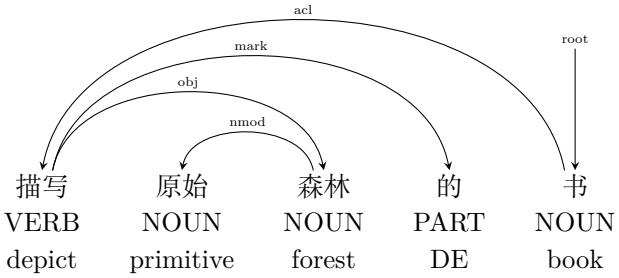
Here the noun is the head, and 的 depends on the adjective with label MARK.



Tree 4: adjective + 的 + noun, from `sent_id=1_1`

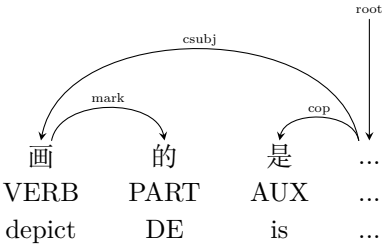
4.2.3 Adjective Clause + 的 + Noun

The structure here is almost the same as in the previous case.



Tree 5: acl + 的 + noun, from `sent_id=1_1`

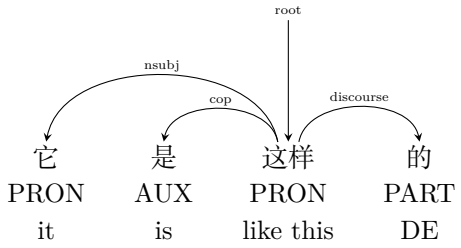
When used as a subject:



Tree 6: acl + 的 as subject, from `sent_id=1_2`

4.2.4 的 at the end of the sentence

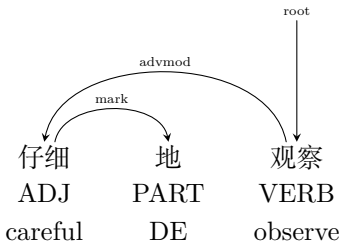
At the end of the sentence, 的 has label DISCOURSE.



Tree 7: end of the sentence 的, from `sent_id=1_2`

4.3 地 (de5)

地 (de5) is used to convert an adjective into an adverb.

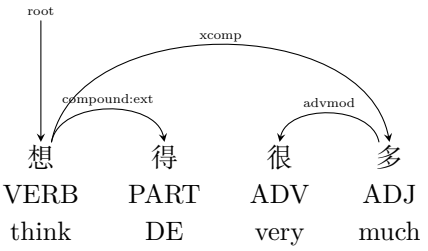


Tree 8: adjective + 地, from `sent_id=1_1`

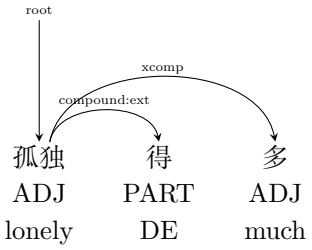
4.4 得 (de5)

得 (de5) is used in two special construction called "descriptive complements" and "complements of extent". They are treated using the same label COMPOUND:EXT.

In the first one, an adjective placed after a verb or an adjective to add a description to the verb or adjective.

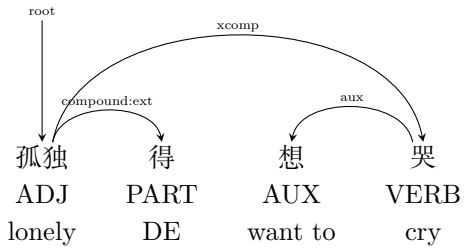


Tree 9: verb + 得, from `sent_id=1_6`



Tree 10: adjective + 得, from `sent_id=2_7`

In the second case, a clause occurs after an adjective to describe the extent of the adjective. One of XCOMP or CCOMP should be used.

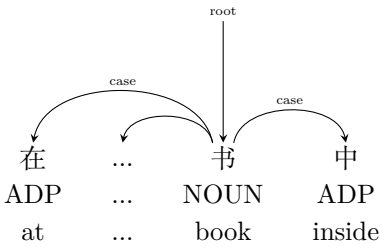


Tree 11: adjective + 得 + extent, modified from `sent_id=2_7`

We note that the PUD treebank does something different. We choose to use this strategy in the UD documentation.

4.5 Adpositions: 在 (zai4), 对 (dui4), 中 (zhong1), 上 (shang4) etc.

A noun can take a preposition or a postposition, or both. In all cases, the adposition has dependency relation CASE. An example is:

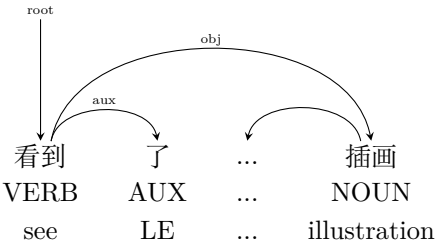


Tree 12: adpositions, from `sent_id=1_1`

We note that currently in the GSD treebank, postpositions are labeled with ACL, which is mostly likely a mistake during the automatic conversion.

4.6 Aspect markers: 了 (le5), 着 (zhe5), 过 (guo4)

These words are given the part-of-speech tag of AUX, and the dependency relation is also AUX.

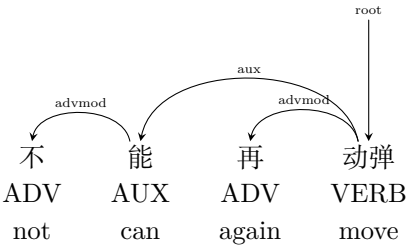


Tree 13: aspect marker, from `sent_id=1_1`

We note that currently in the GSD treebank, aspect markers are labeled with CASE, which is mostly likely a mistake during the automatic conversion.

4.7 Modal verbs: 能 (neng2), 会 (hui4), 没有 (mei2you3), etc.

These words are treated as AUX in the UD, and the dependency relation is also AUX.

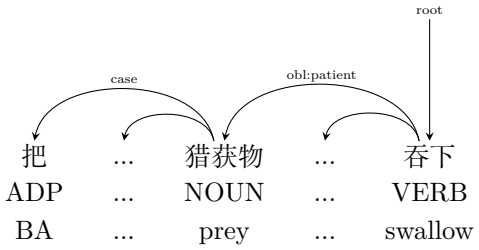


Tree 14: model verb, from `sent_id=1_4`

Note that by UD documentation, 不 is allowed to be the dependent of an AUX.

4.8 把 (ba3)

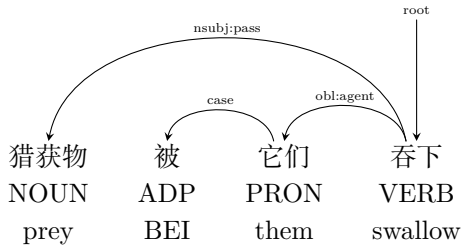
把 puts the object of a verb in front of the verb. UD has a special label for this construction.



Tree 15: 把, from `sent_id=1_4`

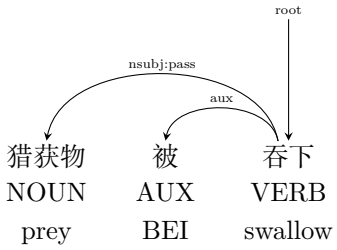
4.9 被 (bei4)

被 is a passive construction. It has a "long" version and a "short" version. In the long version, the agent is present, and it is very similar to the 把 case.



Tree 16: long version of 被, from `sent_id=1_4`, converted into passive

In the short version, the agent is absent, and 被 is consider an AUX of the verb.



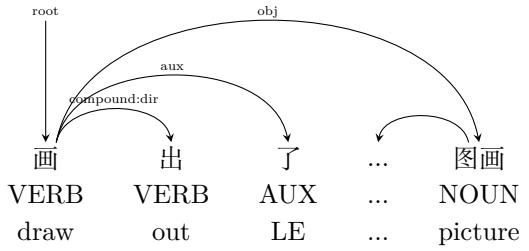
Tree 17: short version of 被, from `sent_id=1_4`, converted into passive

4.10 Compound verbs

In UD, COMPOUND is used for compound nouns. For compound verbs, the subtypes COMPOUND:DIR, COMPOUND:VO, and COMPOUND:VV are used. In all cases, this label connects two words that can be considered as one verb.

4.10.1 Verb + direction

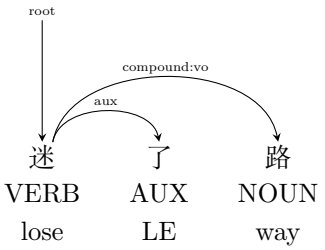
COMPOUND:DIR connects a verb to a direction. For example, 画出 = draw + out = finish drawing.



Tree 18: COMPOUND:DIR, from `sent_id=1_6`

4.10.2 Verb + object

From UD documentation, COMPOUND:VO is for "verb-object compounds where the combination is semantically one unit but syntactically separate". For example, 说话 = say + words = say.



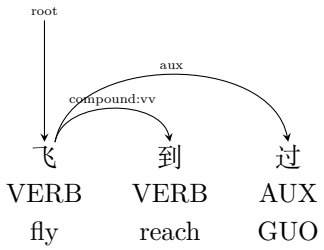
Tree 19: COMPOUND:VO, from `sent_id=2_23`

4.10.3 Verb + verb

COMPOUND:VV is used when (1) the second verb (or occasionally adjective) describes the result of the first verb, or (2) when the second verb is 着 (zhao2), 到 (dao4), 见 (jian4), 完 (wan2), or 过 (guo4). Note that these are different from the auxiliary 着 (zhe5) and 过 (guo4).

4.10.4 Other properties of compound verbs

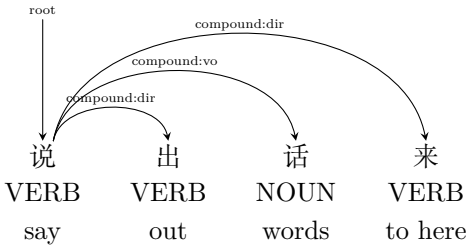
In all cases, the subjects, objects, etc. are dependents of the first verb.



Tree 20: COMPOUND:VV, from `sent_id=1_21`

In COMPOUND:DIR and COMPOUND:VO, the auxiliaries can sometimes go between the two parts (as in tree 19). In COMPOUND:VV, only 不 can go between the two parts.

Multiple compounds can happen at the same time.



Tree 21: multiple compounds, from `sent_id=2_25`

Bibliography

[1] Xia, Fei, *The Segmentation Guidelines for the Penn Chinese Treebank (3.0)* (2000). IRCS Technical Reports Series. 37. `http://repository.upenn.edu/ircs_reports/37`.

Index